「情報量統計学」の３章と４章

AIC入門

３章 推定

４章 AIC

３　推定

この章のトピック ── 最尤法

- エントロピーと情報量
- 情報量の推定値 ── 対数尤度
- 最尤法

モデルの評価，推定

偶然を伴う現象，複雑な現象 ── ある確率分布に従う確率変数の実現値とみなす

(1) データを生成する機構　(2) 現象を表現するモデル

ともに確率分布であるとみなす立場をとる

- 真の確率分布を近似する「モデル」の近似の良さの評価の方法

  モデルが確率分布に対してどの程度近いかを評価する

- 与えられたデータからモデルを推定する方法

  なるべく近似の良いモデル

3.1　エントロピーと情報量

[注意] 教科書の表記とはpとqを入れ替えてある．

m個の事象$\omega_1, \ldots, \omega_m$の確率を$q_1, \ldots, q_m$とする離散分布を考える

$$\boldsymbol{q} = (q_1, \ldots, q_m); \quad q_1 + \cdots + q_m = 1,\ q_i > 0$$

[例] ある球団の勝つ確率

予想Ａ ７割: $\boldsymbol{p}_A = (0.7, 0.3)$
予想Ｂ ５割: $\boldsymbol{p}_B = (0.5, 0.5)$

もし真の勝率が４割なら$\boldsymbol{q} = (0.4, 0.6)$なので，予想Ｂのほうが予想Ａより良い．

もし真の勝率が６割なら$\boldsymbol{q} = (0.6, 0.4)$だが，どちらがよいか？

真の分布とモデルの「近さ」を測る客観的な規準が必要

K-L情報量（相対エントロピー）

真の離散分布 $\boldsymbol{q} = (q_1, \ldots, q_m)$，モデル $\boldsymbol{p} = (p_1, \ldots, p_m)$

$$I(\boldsymbol{q}; \boldsymbol{p}) = \sum_{i=1}^{m} q_i \log \frac{q_i}{p_i}$$

log は e を底とする自然対数．

モデル p に関する真の分布 q の Kullback-Leibler（カルバック・ライブラ）情報量．

K-L情報量を分布の近さを測る規準として採用する
$I(q; p)$ の値が小さく0に近いほどモデル p は真の分布 q に近いとみなす

[例] 勝率予想

$$I(\boldsymbol{q}; \boldsymbol{p}_A) = 0.6 \log \frac{0.6}{0.7} + 0.4 \log \frac{0.4}{0.3} = 0.0226$$

$$I(\boldsymbol{q}; \boldsymbol{p}_B) = 0.6 \log \frac{0.6}{0.5} + 0.4 \log \frac{0.4}{0.5} = 0.0201$$

予想Bの方が（わずかに）良い

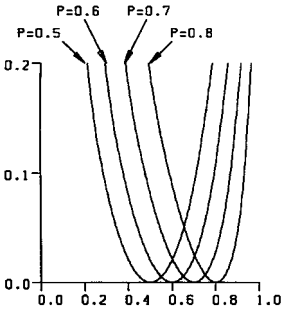
K-L情報量の性質

(i) $I(\boldsymbol{q}; \boldsymbol{p}) \geq 0$

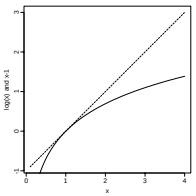
(ii) $I(\boldsymbol{q}; \boldsymbol{p}) = 0 \Leftrightarrow \boldsymbol{p} = \boldsymbol{q} \quad (p_i = q_i,\ i = 1, \ldots, m)$

[証明] $\log x \leq x - 1$ **で等号は** $x = 1$ **のときのみ．** $x = p_i/q_i$ **を代入して**

$$-I(\boldsymbol{q}; \boldsymbol{p}) = \sum_{i=1}^{m} q_i \log \frac{p_i}{q_i} \leq \sum_{i=1}^{m} q_i \left(\frac{p_i}{q_i} - 1 \right) = \sum_{i=1}^{m} (p_i - q_i) = 0$$



$\log(x)$ **と** $x - 1$ **のプロット**



$q = 0.5, 0.6, 0.7, 0.8$ **と変えたときの**
$I((q, 1-q); (p, 1-p))$ **の** p **に対するプロット．**

K-L情報量の解釈

$\boldsymbol{n} = (n_1, \ldots, n_m)$ **が多項分布に従うとする．** $\sum_i n_i = n$ **で確率ベクトルは** $\boldsymbol{p}$ **とする．**

$n\boldsymbol{q}$ **の要素が整数のとき，**

$$\log P\left(\frac{\boldsymbol{n}}{n} = \boldsymbol{q} \mid \boldsymbol{p} \right) \approx -nI(\boldsymbol{q}; \boldsymbol{p})$$

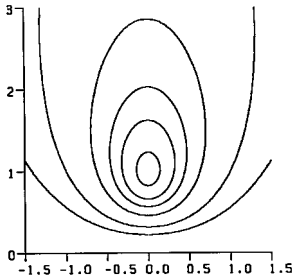
[参考] $-I(q; p)$ **は負のエントロピーと呼ばれる．**

真の密度関数 $g(x)$**，モデル** $f(x)$

$$I(g; f) = \int_{-\infty}^{\infty} g(x) \log \frac{g(x)}{f(x)}\, dx$$

[例] 真の分布 g**:** $N(\mu, \sigma^2)$**，モデル** f**:** $N(\xi, \tau^2)$

$$g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right), \quad f(x) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(x-\xi)^2}{2\tau^2} \right)$$



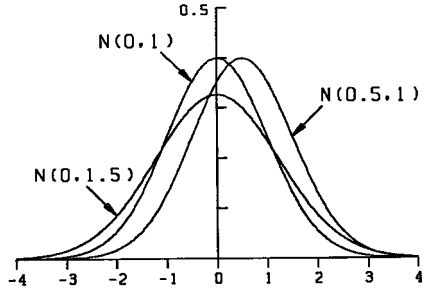
$$I(g; f) = \frac{1}{2} \left\{ \log \frac{\tau^2}{\sigma^2} + \frac{\sigma^2 + (\mu - \xi)^2}{\tau^2} - 1 \right\}$$

$(\mu, \sigma^2) = (0, 1)$ **のときの** $I(g; f)$ **の** (ξ, τ^2) **に対するプロット**

[例] つぎの二つのモデルのうちどちらが真の分布 $N(0,1)$ に近いか？

$f_1 : N(0.5,1)$ **の K-L 情報量は** $I(g;f_1) = 0.125$

$f_2 : N(0,1.5)$ **の K-L 情報量は** $I(g;f_2) = 0.036$



[参考] 情報理論におけるシャノンのエントロピー

$$H(\boldsymbol{q}) = \sum_{i=1}^{m} q_i \log \frac{1}{q_i}$$

対数の底が e **ならば単位は nat（ナット），底が 2 ならば単位は bit（ビット）となる．**

シャノンのエントロピーと K-L 情報量の関係

$$H(\boldsymbol{q}) = H(\boldsymbol{p}_0) - I(\boldsymbol{q};\boldsymbol{p}_0)$$

ただし $\boldsymbol{p}_0 = (1/m,\ldots,1/m)$，$H(\boldsymbol{p}_0) = \log m$．

[参考] 情報理論における相互情報量

$$H(X;Y) = \sum_{x_i}\sum_{y_j} p(x_i,y_j) \log \frac{p(x_i,y_j)}{p(x_i)p(y_j)}$$

相互情報量と K-L 情報量の関係

$$H(X;Y) = I(p(x,y);p(x)p(y))$$

3.2 情報量の推定値 ── 対数尤度

実際の応用では真の分布 q は未知．(連続型の時は $g(\cdot)$ が未知)．

データから K-L 情報量を推定する

[例] ある球団の勝つ確率

予想 A 7 割: $\boldsymbol{p}_A = (0.7, 0.3)$
予想 B 5 割: $\boldsymbol{p}_B = (0.5, 0.5)$

100 試合終った段階での実際の成績は 65 勝 35 敗だった．

どちらの予想が良かったか？

平均対数尤度

K-L 情報量の式を展開すると

$$I(\boldsymbol{q};\boldsymbol{p}) = \sum_{i=1}^{m} q_i \log \frac{q_i}{p_i} = \sum_{i=1}^{m} q_i \log q_i - \sum_{i=1}^{m} q_i \log p_i$$

この第 2 項を大きくするモデル p **が良い**

平均対数尤度 (の n **倍)**

$$l^*(\boldsymbol{p}) = n \sum_{i=1}^{m} q_i \log p_i$$

これを大きくする p **が良いモデル**

対数尤度

事象w_iを観測した回数をn_iとすると，相対頻度n_i/nはq_iの推定になっている．

$l^*(\boldsymbol{p})$のq_iにn_i/nを代入して次の推定値を得る．

対数尤度

$$l(\boldsymbol{p}) = \sum_{i=1}^{m} n_i \log p_i$$

これを大きくするpは良いモデルの推定になる．

$i = 1, \ldots, n$に対してi番目の観測値をx_iとすると

$$l(\boldsymbol{p}) = \sum_{i=1}^{n} \log p_{x_i}$$

とも書ける．

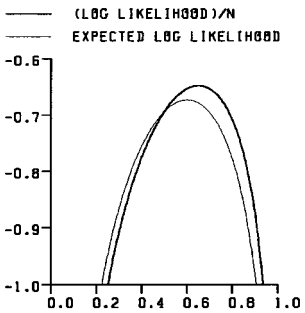
[例] 球団の勝率予想 $n_1 = 65, n_2 = 35$

$$l(\boldsymbol{p}) = 65 \log p_1 + 35 \log p_2$$

$$l(\boldsymbol{p}_A) = 65 \log 0.7 + 35 \log 0.3 = -65.32$$

$$l(\boldsymbol{p}_B) = 65 \log 0.5 + 35 \log 0.5 = -69.31$$

データからは予想Ａの方が良かったと判断される．



$n \to \infty$の極限では，

$$\frac{1}{n} l(\boldsymbol{p}) \to \frac{1}{n} l^*(\boldsymbol{p})$$

であるが有限のnでは多少ずれる．

連続型確率変数の場合

K-L情報量

$$\begin{aligned} I(g; f) &= \int_{-\infty}^{\infty} g(x) \log \frac{g(x)}{f(x)}\, dx \\ &= \int_{-\infty}^{\infty} g(x) \log g(x)\, dx - \int_{-\infty}^{\infty} g(x) \log f(x)\, dx \end{aligned}$$

平均対数尤度(のn倍)

$$l^*(f) = n \int_{-\infty}^{\infty} g(x) \log f(x)\, dx$$

対数尤度

$$l(f) = \sum_{i=1}^{n} \log f(x_i)$$

$n \to \infty$の極限では

$$\frac{1}{n} l(f) \to \frac{1}{n} l^*(f)$$

[例] 10個の観測値 {-1.10，-0.40，-0.20，-0.02，0.02，0.71，1.35，1.46，1.74，3.89}**が得られたとき，つぎの二つのモデルのうちどちらが良いか？**

$$f_1(x) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right) \quad \text{正規分布}$$

$$f_2(x) = \frac{1}{\pi(x^2+1)} \quad \text{コーシー分布}$$

$$\sum_{i=1}^{10} \log f_1(x_i) = -21.20, \quad \sum_{i=1}^{10} \log f_2(x_i) = -19.19$$

コーシー分布の方が良い．

もし最後の観測値3.89（外れ値？）をとりのぞくと，

$$\sum_{i=1}^{9} \log f_1(x_i) = -12.72, \quad \sum_{i=1}^{9} \log f_2(x_i) = -15.26$$

正規分布の方が良い．

尤度

確率変数 $(X_1, \ldots, X_n)$ のパラメトリック・モデル

$$f(x_1, \ldots, x_n|\boldsymbol{\theta})$$

データ $(x_1, \ldots, x_n)$ を観測したときの尤度 (likelihood)

$$L(\boldsymbol{\theta}) = f(x_1, \ldots, x_n|\boldsymbol{\theta})$$

対数尤度は，尤度の対数として定義される

$$l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$$

$X_1, \ldots, X_n$ **が独立なら，**

$$L(\boldsymbol{\theta}) = f(x_1|\boldsymbol{\theta}) \cdots f(x_n|\boldsymbol{\theta}) \quad \textbf{or} \quad p(x_1|\boldsymbol{\theta}) \cdots p(x_n|\boldsymbol{\theta})$$

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(x_i|\boldsymbol{\theta}) \quad \textbf{or} \quad \sum_{i=1}^{n} \log p(x_i|\boldsymbol{\theta})$$

3.3 最尤法

パラメトリック・モデルにおいて，データが与えられたとき，尤度 $L(\theta)$ もしくは対数尤度 $l(\theta)$ を最大とするようなパラメタの値 $\hat{\theta}$ を選ぶ
このパラメタ推定法は，最大尤度法，略して「最尤法」と呼ばれる．

推定されたパラメタ値 $\hat{\theta}$ は，最尤推定量．

$$l(\theta) \text{ を最大にする } \theta \Rightarrow \hat{\theta}$$

これにより決まるモデルは，最尤モデル

$$f(x_1, \ldots, x_n|\hat{\boldsymbol{\theta}}) \quad \textbf{or} \quad p(x_1, \ldots, x_n|\hat{\boldsymbol{\theta}})$$

最尤モデルの対数尤度は，最大対数尤度

$$l(\hat{\boldsymbol{\theta}}) = \max_{\boldsymbol{\theta} \in \Theta} l(\boldsymbol{\theta})$$

最尤法 ── 2項分布

$$p(k|\theta) = \begin{pmatrix} n \\ k \end{pmatrix} \theta^k (1-\theta)^{n-k}; \quad k = 0, 1, \ldots, n$$

$$l(\theta) = \log \begin{pmatrix} n \\ k \end{pmatrix} + k \log \theta + (n-k) \log(1-\theta)$$

これを最大にする θ を求めるために θ で微分すると

$$\frac{dl}{d\theta} = \frac{k}{\theta} - \frac{n-k}{1-\theta}$$

これを 0 とおいて，

$$\hat{\theta} = \frac{k}{n}$$

最尤法 ── 多項分布

c 個の事象の起こった度数を $n_1, \ldots, n_c$，$n_1 + \cdots + n_c = n$ とする．

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{c} n_i \log \theta_i$$

$\sum_{i=1}^{c} \theta_i = 1$ **の条件の下で** $l(\theta)$ **を最大にする** $\hat{\boldsymbol{\theta}}$ **は，**

$$\hat{\theta}_i = \frac{n_i}{n}$$

として得られる．

$$l(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{c} n_i \log n_i - n \log n$$

最尤法 ── ポアソン分布

$$\log p(k|\lambda) = -\lambda + k\log\lambda - \log k!$$

i番目の観測値をk_iとする．

$$l(\lambda) = \sum_{i=1}^{n} \log p(k_i|\lambda) = \sum_{i=1}^{n} (-\lambda + k_i \log\lambda - \log k_i!)$$

$$\frac{dl(\lambda)}{d\lambda} = \sum_{i=1}^{n} \left(-1 + \frac{k_1}{\lambda}\right)$$

これを0とおいて

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} k_i$$

最尤法 ── 正規分布 $N(\mu, \sigma^2)$

$$l(\mu, \sigma^2) = -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\begin{aligned}
\frac{\partial l}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0 \\
\frac{\partial l}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 = 0
\end{aligned}$$

を解いて，

$$\begin{aligned}
\hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2
\end{aligned}$$

これを$l(\mu, \sigma^2)$に代入すると

$$l(\hat{\mu}, \hat{\sigma}^2) = -\frac{n}{2}\log 2\pi\hat{\sigma}^2 - \frac{n}{2}$$

最尤法の性質（モデルが正しいとき）

最尤推定量の漸近(ぜんきん)正規性
$X_1, \ldots, X_n \sim f(x|\boldsymbol{\theta}^*)$ **なら十分大きなnに対して近似的に**

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}^*, G^{-1}/n)$$

$$G_{ij} = E_X\left[\frac{\partial \log f(X|\boldsymbol{\theta})}{\partial\theta_i}\frac{\partial \log f(X|\boldsymbol{\theta})}{\partial\theta_j}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$$

$n \to \infty$のとき

- $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}^*$ **(一致性)**
- $E(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}^* + O(n^{-1})$ **(漸近不偏性)**
- **他の推定量ではG^{-1}/nより小さな分散には出来ない（漸近有効性)**

最尤法の性質

$X_1, \ldots, X_n \sim g(x)$ **なら十分大きなnに対して近似的に**

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}^*, V/n)$$

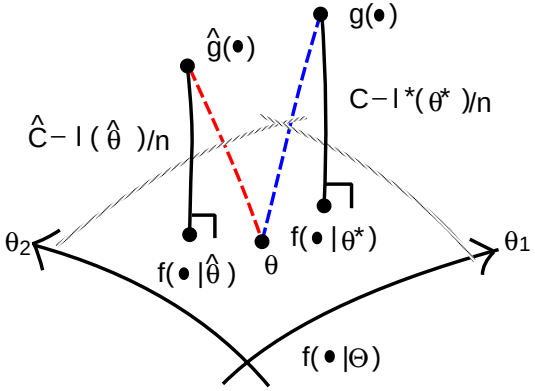
$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\Theta} I(g; f(\boldsymbol{\theta})), \quad V = H^{-1}GH^{-1}$$

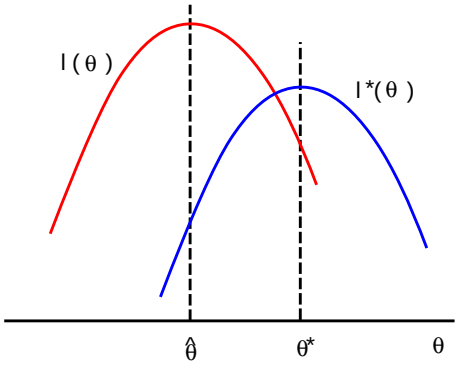
$$G_{ij} = E_X\left[\frac{\partial \log f(X|\boldsymbol{\theta})}{\partial\theta_i}\frac{\partial \log f(X|\boldsymbol{\theta})}{\partial\theta_j}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}, H_{ij} = -E_X\left[\frac{\partial^2 \log f(X|\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$$

$f(\cdot|\boldsymbol{\theta}^*)$はパラメトリック・モデル$f(\cdot|\Theta)$のなかで最も$g(\cdot)$に近い分布．

パラメトリック・モデルが正しい場合 $\Leftrightarrow$ **$f(\cdot|\Theta)$が$g(\cdot)$を含んでいる**

$$g(\cdot) = f(\cdot|\boldsymbol{\theta}^*), \quad H = G = \text{Fisher情報行列なので } V = G^{-1} = H^{-1}$$





$$
\begin{aligned}
I(g; f(\boldsymbol{\theta})) &= \int g(x) \log g(x)\, dx - \int g(x) \log f(x|\boldsymbol{\theta})\, dx \\
&= C - l^*(\boldsymbol{\theta})/n
\end{aligned}
$$

$$
\begin{aligned}
I(\hat{g}; f(\boldsymbol{\theta})) &= \int \hat{g}(x) \log \hat{g}(x)\, dx - \int \hat{g}(x) \log f(x|\boldsymbol{\theta})\, dx \\
&= \hat{C} - l(\boldsymbol{\theta})/n
\end{aligned}
$$

最尤推定量の幾何的解釈

経験分布

$$\hat{g}(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x_i = x)$$

対数尤度

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(x_i|\boldsymbol{\theta}) = n \int \hat{g}(x) \log f(x|\boldsymbol{\theta})\, dx$$

K-L 情報量

$$
\begin{aligned}
I(\hat{g}; f(\boldsymbol{\theta})) &= \int \hat{g}(x) \log \hat{g}(x)\, dx - \int \hat{g}(x) \log f(x|\boldsymbol{\theta})\, dx \\
&= \hat{C} - l(\boldsymbol{\theta})/n
\end{aligned}
$$

(第１項は連続型の場合は発散してしまう．)

- $\hat{\boldsymbol{\theta}}$ は $I(\hat{g}; f(\boldsymbol{\theta}))$ を最小にするパラメタ値．
- $f(\cdot|\hat{\boldsymbol{\theta}})$ は $f(\cdot|\Theta)$ のなかで最も $g(\cdot)$ に近い分布．

最尤推定量の漸近分布の導出

$$\left[\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]_{\theta=\hat{\theta}} = \sum_{i=1}^{n} \left[\frac{\partial \log f(x_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]_{\theta=\hat{\theta}} = 0$$

これを θ^* の周辺でテーラ展開して

$$\left[\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]_{\theta=\hat{\theta}} \approx \left[\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]_{\theta=\theta^*} + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \left[\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}' \partial \boldsymbol{\theta}}\right]_{\theta=\theta^*}$$

中心極限定理より

$$\frac{1}{\sqrt{n}} \left[\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]_{\theta=\theta^*} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\frac{\partial \log f(x_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]_{\theta=\theta^*} \sim N(0, G)$$

大数の法則より

$$\frac{1}{n} \left[\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}' \partial \boldsymbol{\theta}}\right]_{\theta=\theta^*} \to -H$$

したがって

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \approx H^{-1} \frac{1}{\sqrt{n}} \left[\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]_{\theta=\theta^*} \sim N(0, H^{-1} G H^{-1})$$

$g(\cdot) = f(\cdot|\theta)$ のときに $H = G$ となることの証明

$\log f$ を θ_i で微分

$$\frac{\partial \log f}{\partial \theta_i} = \frac{1}{f} \frac{\partial f}{\partial \theta_i}$$

もう一度 θ_j で微分

$$\frac{\partial^2 \log f}{\partial \theta_i \partial \theta_j} = -\frac{1}{f^2} \frac{\partial f}{\partial \theta_i} \frac{\partial f}{\partial \theta_j} + \frac{1}{f} \frac{\partial^2 f}{\partial \theta_i \partial \theta_j}$$

したがって，

$$\frac{\partial^2 \log f}{\partial \theta_i \partial \theta_j} = -\frac{\partial \log f}{\partial \theta_i} \frac{\partial \log f}{\partial \theta_j} + \frac{1}{f} \frac{\partial^2 f}{\partial \theta_i \partial \theta_j}$$

両辺を $f(x|\theta)$ に関する期待値をとると，

$$-H_{ij}(\boldsymbol{\theta}) = -G_{ij}(\boldsymbol{\theta}) + \int \frac{\partial^2 f(x|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\, dx$$

最後の項は０

4 AIC

この章のトピック ── 赤池情報量規準 AIC

- **概要**
- **期待平均対数尤度**
- **AICの導出**

モデル選択

複数のパラメトリック・モデルから，良いものを選ぶ．

$$\mathrm{AIC} = -2 \times \textbf{(最大対数尤度)} + 2 \times \textbf{(自由パラメタ数)}$$

をモデル選択の規準とする

候補となるモデル毎にAICを計算して，その値を最小にするモデルを選択する．
このモデルを，最小AIC推定値 (MAICE) とよぶ

- **AICの第１項: モデルのデータへの当てはまりの良さ**
- **AICの第２項: モデルの複雑さ**

同程度にあてはまりの良いモデルが複数ある場合，そのなかで自由パラメタ数の小さいものを選ぶべきである ── 節約の原理

4.1 概要

パラメトリック・モデルの良さの規準

$$\textbf{期待平均対数尤度} := E\Big\{\textbf{平均対数尤度}\Big\}$$

期待平均対数尤度を最大対数尤度で推定 ── 偏りがある

偏りを補正

$$E\Big\{\textbf{最大対数尤度} - \textbf{自由パラメタ数}\Big\} \approx \textbf{期待平均対数尤度}$$

$$\mathrm{AIC} := -2 \times \Big(\textbf{最大対数尤度} - \textbf{自由パラメタ数}\Big)$$

定義と仮定

確率変数 $\boldsymbol{X} = (X_1, \ldots, X_n)$**，データは実現値** $\boldsymbol{x} = (x_1, \ldots, x_n)$**．**

真の分布

$$X_1, \ldots, X_n \sim g(\cdot) \qquad \textbf{独立に同分布に従う}$$

$$g(x_1, \ldots, x_n) = g(x_1) \cdots g(x_n)$$

モデル MODEL(K)

$$X_1, \ldots, X_n \sim f(\cdot|\boldsymbol{\theta}) \qquad \textbf{独立に同分布に従う}$$

$$f(x_1, \ldots, x_n|\boldsymbol{\theta}) = f(x_1|\boldsymbol{\theta}) \cdots f(x_n|\boldsymbol{\theta})$$

対数尤度

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(x_i|\boldsymbol{\theta}); \quad \boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$$

MODEL(K)に制約を入れた，部分モデル MODEL(k)

$$\Theta_k = \{\boldsymbol{\theta} \in \Theta_K | \theta_{k+1} = \theta_{k+2} = \cdots = \theta_K = 0\}$$

MODEL(k)の自由パラメタ数 $= k$

$$\Theta_1 \subset \Theta_2 \subset \cdots \subset \Theta_K$$

データxが与えられたときのMODEL(k)のパラメタの最尤推定量を$\hat{\theta}_k$とする．

最大対数尤度

$$l(\hat{\boldsymbol{\theta}}_k) = \max_{\theta \in \Theta_k} l(\boldsymbol{\theta})$$

情報量規準

$$\mathrm{AIC}(k) = -2 \times \left(l(\hat{\boldsymbol{\theta}}_k) - k\right)$$

4.2 期待平均対数尤度

分布$f(\cdot|\theta)$の平均対数尤度は，

$$E_Z\{\log f(Z|\boldsymbol{\theta})\} = \int g(z) \log f(z|\boldsymbol{\theta})\, dz; \qquad Z \sim g(\cdot)$$

これをn倍して

$$l^*(\boldsymbol{\theta}) = nE_Z\{\log f(Z|\boldsymbol{\theta})\}$$

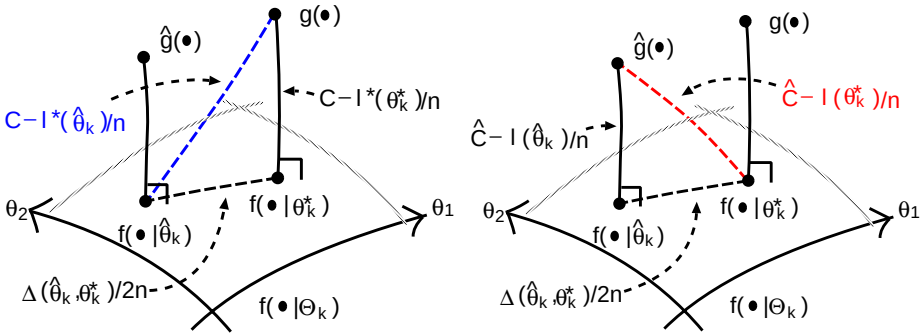
$l^*(\theta)$が大きい程$f(\cdot|\theta)$の$g(\cdot)$に対する近似が良いことになる．

MODEL(k)の最尤モデルの良さは$f(\cdot|\hat{\theta}_k)$の平均対数尤度$l^*(\hat{\theta}_k)$で評価

$l^*(\hat{\theta}_k)$のデータに関する期待値を取って「期待平均対数尤度」

$$l_n^*(k) = E_X\{l^*(\hat{\boldsymbol{\theta}}_k)\} = \int l^*(\hat{\boldsymbol{\theta}}_k) g(\boldsymbol{x})\, d\boldsymbol{x}$$

でパラメトリック・モデル MODEL(k) を評価する



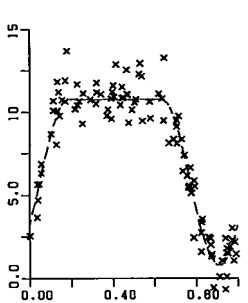
$$E_X\left\{I(g; f(\hat{\boldsymbol{\theta}}_k))\right\} = C - l_n^*(k)/n$$

$$I(g; f(\hat{\boldsymbol{\theta}}_k)) - I(g; f(\boldsymbol{\theta}_k^*)) \approx \Delta(\hat{\boldsymbol{\theta}}_k, \boldsymbol{\theta}_k^*)/2n$$

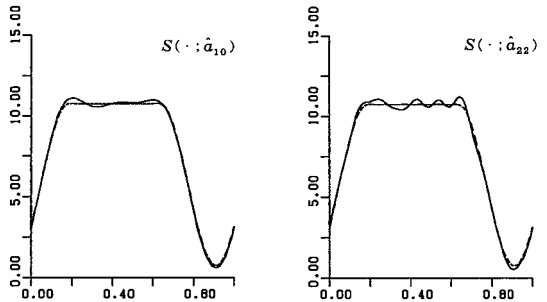
$$I(\hat{g}; f(\boldsymbol{\theta}_k^*)) - I(\hat{g}; f(\hat{\boldsymbol{\theta}}_k)) \approx \Delta(\hat{\boldsymbol{\theta}}_k, \boldsymbol{\theta}_k^*)/2n$$

$$E_X\{l(\boldsymbol{\theta}_k^*)\} = l^*(\boldsymbol{\theta}_k^*)$$

[数値例] 条件つき分布モデル（フーリエ級数による回帰モデル）



$K = 22$のモデルから
データを生成



$k = 10$と$k = 22$のモデルの当てはめ

MODEL(K):

$$Y \sim N(S(x;\boldsymbol{a}),\sigma^2), \quad X \sim [0,1]\text{の一様分布}$$

$$S(x;\boldsymbol{a}) = a_0 + \sum_{m=1}^{M} (a_{2m-1}\sin 2m\pi x + a_{2m}\cos 2m\pi x)$$

$$f(x,y|\boldsymbol{\theta}) = \begin{cases} \exp(-(y-S(x;\boldsymbol{a}))^2/2\sigma^2)/\sqrt{2\pi\sigma^2} & 0 \le x \le 1 \\ 0 & x < 0,\ x > 1 \end{cases}$$

$$\boldsymbol{\theta} \in \Theta = \{(\theta_1,\ldots,\theta_K) = (\sigma^2, a_0, \ldots, a_{2M}) | \sigma^2 > 0\}$$

$$K = 2M + 2$$

$$\Theta_1 \subset \Theta_2 \subset \cdots \subset \Theta_K$$

表4.1の最後の2行

$$l(\hat{\boldsymbol{\theta}}_2) \le l(\hat{\boldsymbol{\theta}}_4) \le \cdots \le l(\hat{\boldsymbol{\theta}}_K)$$

最大対数尤度 $l(\hat{\boldsymbol{\theta}}_k)$は$k$とともに大きくなる

$$l^*_{500}(2) \le l^*_{500}(4) \le \cdots \le l^*_{500}(10) \ge l^*_{500}(12) \ge \cdots \ge l^*_{500}(22)$$

期待平均対数尤度 $l^*_n(k)$は$k = 10$で最大値をとる

真のモデルは$k = 22$は必ずしも期待平均対数尤度を大きくしない

表4.1　最尤推定値，最大対数尤度および期待平均対数尤度

θ	$\hat{\theta}_2$	$\hat{\theta}_4$	$\hat{\theta}_6$	$\hat{\theta}_8$	$\hat{\theta}_{10}$	$\hat{\theta}_{12}$	$\hat{\theta}_{14}$	$\hat{\theta}_{16}$	$\hat{\theta}_{18}$	$\hat{\theta}_{20}$	$\hat{\theta}_{22}$	θ^*
θ_1	14.220	4.145	1.121	0.984	0.974	0.970	0.968	0.966	0.956	0.955	0.949	1.000
θ_2	8.012	7.983	8.043	8.010	8.010	8.012	8.013	8.013	8.014	8.013	8.011	8.000
θ_3		2.470	2.491	2.488	2.484	2.484	2.479	2.479	2.470	2.468	2.465	2.415
θ_4		−3.684	−3.869	−3.854	−3.850	−3.851	−3.853	−3.852	−3.847	−3.847	−3.847	−3.806
θ_5			2.252	2.238	2.242	2.248	2.248	2.247	2.247	2.253	2.249	2.119
θ_6			−0.984	−0.987	−0.987	−0.995	−0.994	−0.997	−1.005	−1.006	−1.009	−0.997
θ_7				0.523	0.516	0.515	0.518	0.518	0.512	0.515	0.515	0.545
θ_8				−0.022	−0.013	−0.012	−0.010	−0.007	0.002	0.004	0.007	0.069
θ_9					−0.129	−0.128	−0.130	−0.126	−0.133	−0.136	−0.139	−0.094
θ_{10}					−0.074	−0.074	−0.074	−0.073	−0.072	−0.071	−0.073	−0.078
θ_{11}						0.070	0.076	0.079	0.079	0.081	0.089	−0.021
θ_{12}						0.041	0.041	0.038	0.024	0.024	0.021	−0.065
θ_{13}							−0.058	−0.058	−0.061	−0.061	−0.060	−0.011
θ_{14}							0.032	0.036	0.042	0.047	0.044	0.042
θ_{15}								−0.065	−0.067	−0.069	−0.075	−0.011
θ_{16}								0.003	−0.010	−0.010	−0.011	0.010
θ_{17}									0.145	0.146	0.142	0.020
θ_{18}									−0.024	−0.024	−0.016	−0.004
θ_{19}										−0.034	−0.035	0.002
θ_{20}										0.041	0.039	0.001
θ_{21}											0.050	−0.006
θ_{22}											−0.089	−0.009
$-l(\hat{\theta}_k)$	1373.14	1064.97	738.13	705.46	702.93	701.91	701.41	700.90	698.35	697.84	696.30	
$-l^*_{500}(k)$	1371.11	1051.87	750.07	716.43	715.73	716.28	717.20	718.29	719.39	720.57	721.80	

数値例における最大対数尤度と期待平均対数尤度の関係

表 4.2　1000回の実験で求めた最大対数尤度の平均値および$l^*_{500}(k)$との差

	2	4	6	8	10	12
$l(\hat{\theta}_k)$ の平均	−1368.85	−1047.52	−744.53	−708.55	−705.61	−703.94
$l(\hat{\theta}_k) - l^*_{500}(k)$ の平均	**2.26**	**4.35**	**5.54**	**7.88**	**10.12**	**12.34**
	14	**16**	**18**	**20**	**22**	
$l(\hat{\theta}_k)$の平均	−702.68	−701.65	−700.61	−699.55	−698.49	
$l(\hat{\theta}_k) - l^*_{500}(k)$ の平均	**14.52**	**16.64**	**18.75**	**21.02**	**23.31**	

$$E\left\{l(\hat{\boldsymbol{\theta}}_k) - l^*_n(k)\right\} \approx k$$

4.3 AICの導出

[仮定] 以下の議論では真の分布を

$$g(\cdot) = f(\cdot|\boldsymbol{\theta}^*), \quad \boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_K^*)$$

とおく．$\boldsymbol{\theta}^* \in \Theta_k$となる最小の$k$を真の自由パラメタ数$k^*$とする．

$$\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_k^*, 0, \ldots, 0)$$

[情報量規準]

$$\mathrm{AIC}(k) = -2 \times \left(l(\hat{\boldsymbol{\theta}}_k) - k \right)$$

[導出する式]

$$E\left\{ l(\hat{\boldsymbol{\theta}}_k) - k \right\} \approx l_n^*(k)$$

第１段階

平均対数尤度 (のn倍)

$$l^*(\boldsymbol{\theta}) = nE_Z\{\log f(Z|\boldsymbol{\theta})\}$$

をθ^*の周辺でテーラ展開して近似式

$$\begin{aligned} l^*(\boldsymbol{\theta}) &\approx l^*(\boldsymbol{\theta}^*) + n(\boldsymbol{\theta} - \boldsymbol{\theta}^*)E_Z\left[\frac{\partial \log f(Z|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right]_{\theta^*} \\ &\quad + \frac{1}{2}n(\boldsymbol{\theta} - \boldsymbol{\theta}^*)E_Z\left[\frac{\partial^2 \log f(Z|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}' \partial \boldsymbol{\theta}}\right]_{\theta^*}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)' \\ &= l^*(\boldsymbol{\theta}^*) - \frac{1}{2}\sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)H\sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)' \\ &= l^*(\boldsymbol{\theta}^*) - \frac{1}{2}\Delta(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \end{aligned}$$

ただし

$$\Delta(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)H\sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)'$$

MODEL(K)での最尤推定$\hat{\theta}_K$を$l^*(\theta)$に代入すると

$$l^*(\hat{\boldsymbol{\theta}}_K) \approx l^*(\boldsymbol{\theta}^*) - \frac{1}{2}\Delta(\hat{\boldsymbol{\theta}}_K, \boldsymbol{\theta}^*)$$

ただし

$$\hat{\boldsymbol{\theta}}_K \sim N(\boldsymbol{\theta}^*, V/n)$$

ここで$H = G$，$V = H^{-1}GH^{-1} = H^{-1}$を利用すると，

$$\Delta(\hat{\boldsymbol{\theta}}_K, \boldsymbol{\theta}^*) \sim \chi_K^2$$

$E\{\chi_K^2\} = K$**なので**

$$l_n^*(K) = E_X\{l^*(\hat{\boldsymbol{\theta}}_K)\} \approx l^*(\boldsymbol{\theta}^*) - \frac{K}{2}$$

MODEL(k)においては$H_k \approx G_k$がいえて，やはり

$$l_n^*(k) = E_X\{l^*(\hat{\boldsymbol{\theta}}_k)\} \approx l^*(\boldsymbol{\theta}_k^*) - \frac{k}{2}$$

もうすこし厳密にやると，

$$\begin{aligned} l_n^*(k) &\approx l^*(\boldsymbol{\theta}_k^*) - \frac{1}{2}\mathrm{tr}\left(H_k E_X\left\{ n(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*)'(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*) \right\} \right) \\ &\approx l^*(\boldsymbol{\theta}_k^*) - \frac{1}{2}\mathrm{tr}\left(H_k H_k^{-1} G_k H_k^{-1} \right) \\ &= l^*(\boldsymbol{\theta}_k^*) - \frac{1}{2}\mathrm{tr}\left(G_k H_k^{-1} \right) \end{aligned}$$

$$\mathrm{tr}\left(G_k H_k^{-1} \right) \approx k$$

$$G_k = E_X\left[\frac{\partial \log f(X|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\frac{\partial \log f(X|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_k^*} \qquad H_k = -E_X\left[\frac{\partial^2 \log f(X|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}' \partial \boldsymbol{\theta}}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_k^*}$$

ただしここでの計算ではθは最初のk個の要素に制限してある

第２段階

対数尤度

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(x_i|\boldsymbol{\theta})$$

を $\hat{\theta}_K$ **の周辺でテーラ展開すると，**

$$\begin{aligned} l(\boldsymbol{\theta}) &\approx l(\hat{\boldsymbol{\theta}}_K) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_K) \left[\frac{\partial l}{\partial \boldsymbol{\theta}'}\right]_{\hat{\theta}_K} \\ &\quad + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_K) \left[\frac{\partial^2 l}{\partial \boldsymbol{\theta}' \partial \boldsymbol{\theta}}\right]_{\hat{\theta}_K} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_K)' \end{aligned}$$

ここで

$$\frac{1}{n}\left[\frac{\partial^2 l}{\partial \boldsymbol{\theta}' \partial \boldsymbol{\theta}}\right]_{\hat{\theta}_K} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial^2 \log f(x_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}' \partial \boldsymbol{\theta}}\right]_{\hat{\theta}_K} \to E_Z\left[\frac{\partial^2 \log f(Z|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}' \partial \boldsymbol{\theta}}\right]_{\theta^*} = -H$$

を代入すると

$$l(\boldsymbol{\theta}) \approx l(\hat{\boldsymbol{\theta}}_K) - \frac{1}{2}\Delta(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_K)$$

MODEL(K**)での最適なパラメタ** θ^* **を** $l(\theta)$ **に代入すると**

$$l(\boldsymbol{\theta}^*) \approx l(\hat{\boldsymbol{\theta}}_K) - \frac{1}{2}\Delta(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}_K)$$

これに次の二つの式を代入して

$$E_X\{l(\boldsymbol{\theta}^*)\} = E_X\Big\{\sum_{i=1}^{n} \log f(X_i|\boldsymbol{\theta}^*)\Big\} = nE_Z\{\log f(Z|\boldsymbol{\theta}^*)\} = l^*(\boldsymbol{\theta}^*)$$

$$E_X\{\Delta(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}_K)\} \approx K$$

MODEL(K**)では**

$$l^*(\boldsymbol{\theta}^*) \approx E_X\{l(\hat{\boldsymbol{\theta}}_K)\} - \frac{K}{2}$$

同様にMODEL(k**)では**

$$l^*(\boldsymbol{\theta}_k^*) \approx E_X\{l(\hat{\boldsymbol{\theta}}_k)\} - \frac{k}{2}$$

第１段階と第２段階をまとめる

$$\begin{aligned} l_n^*(k) &\approx l^*(\boldsymbol{\theta}_k^*) - \frac{k}{2} \\ l^*(\boldsymbol{\theta}_k^*) &\approx E_X\{l(\hat{\boldsymbol{\theta}}_k)\} - \frac{k}{2} \end{aligned}$$

MODEL(k**)**

$$l_n^*(k) \approx E_X\{l(\hat{\boldsymbol{\theta}}_k)\} - k$$

もうすこし厳密にやると，

$$l_n^*(k) \approx E_X\{l(\hat{\boldsymbol{\theta}}_k)\} - \mathrm{tr}\left(G_k H_k^{-1}\right)$$

赤池の情報量規準

$$\mathrm{AIC}(k) = -2 \times \left(l(\hat{\boldsymbol{\theta}}_k) - k\right)$$

竹内の情報量規準

$$\mathrm{TIC}(k) = -2 \times \left(l(\hat{\boldsymbol{\theta}}_k) - \mathrm{tr}\left(G_k H_k^{-1}\right)\right)$$

AICのバラツキ

表 4.6　AIC最小となる k **(=**k_A**) の分布 (**$n = 500$**)**

k_A	2	4	6	8	10	12	14	16	18	20	22	計
度数	0	0	0	326	342	173	91	27	17	14	10	1000

表 4.7　**$n = 2000$のときの** k_A**の分布**

k_A	2	4	6	8	10	12	14	16	18	20	22	計
度数	0	0	0	6	210	359	264	68	58	21	14	1000

なお，$l_{500}^*(k)$ **最小は** $k = 10$**，**$l_{2000}^*(k)$ **最小は** $k = 12$**のとき**

AIC利用上の注意

- データ数に対して自由パラメタ数が大きすぎる場合はAICは使えない
  (導出の際に，kを固定して$n \to \infty$という漸近論を用いている)
- AICの値そのものよりも，AICの値の差に意味がある
- モデル比較の際は，AICの差のバラツキを考慮する
  とくに「ノンネスト」の場合は，バラツキが非常に大きい！
- モデル比較でAICの値がほぼ等しくても推定された分布の形が大幅に違うこともある
- モデルの次数を上げるほどAICが下がってなかなか最小値を取らないことがある
- AICは真の自由パラメタ数を推定するための規準ではない！